Let's Talk About X: Combining Image Recognition and Eye Gaze to Support Conversation for People with ALS

Shaun K. Kane
University of Colorado Boulder
Boulder, CO, USA
shaun.kane@colorado.edu

Meredith Ringel Morris
Microsoft Research
Redmond, WA, USA
merrie@microsoft.com

ABSTRACT
Communicating at a natural speed is a significant challenge for users of augmentative and alternative communication (AAC) devices, especially when input is provided by eye gaze, as is common for people with ALS and similar conditions. One way to improve AAC throughput is by drawing on contextual information from the outside world. Toward this goal, we present SceneTalk, a prototype gaze-based AAC system that uses computer vision to identify objects in the user's field of view and suggests words and phrases related to the current scene. We conducted a formative evaluation of SceneTalk with six people with ALS, in which we evaluated their preference for user interface modes and output preferences. Participants agreed that integrating contextual awareness into their AAC device could be helpful across a diverse range of situations.

Author Keywords
Assistive technology; eye gaze; computer vision; augmentative and alternative communication; ALS.

ACM Classification Keywords
K.4.2. Social Issues: Assistive technologies for persons with disabilities; H.5.2. User Interfaces: Voice I/O.

INTRODUCTION
Augmentative and alternative communication (AAC) systems provide communication support for people with a range of physical, cognitive, and other disabilities [5]. Designing an effective AAC system requires balancing the user's expressive capability with their ability to quickly and reliably operate the device's user interface.

Creating effective AAC devices for people with neurodegenerative diseases such as Amyotrophic Lateral Sclerosis (ALS) is especially challenging, as these individuals typically retain their full linguistic abilities but often lose the ability to speak and use their hands entirely [2].

DIS 2017, June 10-14, 2017, Edinburgh, United Kingdom

ACM 978-1-4503-4922-2/17/06 $15.00
DOI: http://dx.doi.org/10.1145/3064663.3064762

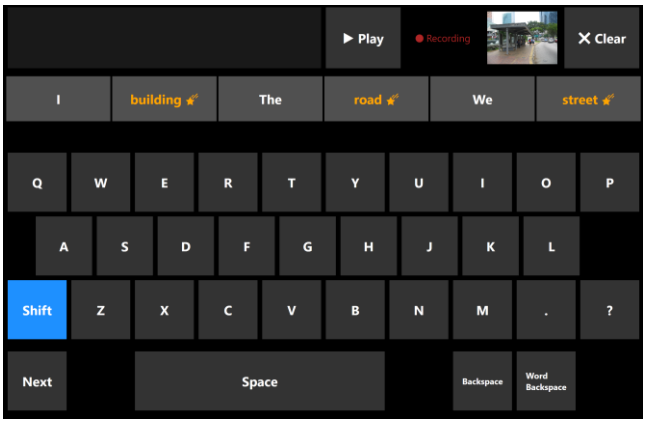

Figure 1. SceneTalk is a gaze-based AAC system that uses the device's camera to recognize objects. Recognized objects are used to suggest relevant words and phrases as the user types. The prediction bar at the top of the keyboard combines contextual predictions detected in the image (in orange) with predictions from a language model (in white).

Many people with ALS use eye gaze to interact with their AAC systems [4]. Using gaze for real-time communication can be challenging, as eye-typing typically enables users to communicate at 10 to 20 words per minute (wpm) [18], while spoken conversation occurs at about 180 wpm [28]. Thus, improving the speed of AAC input has the potential to significantly improve quality of life for people with ALS.

One way to increase the speed of AAC input is to predict likely words and phrases as the user types. Predictions can be generated from a language model (e.g., [23]) or derived from knowledge of the user's context, such as their location (e.g., [7,12,26]). However, AAC systems may be able to infer context from other sources. For example, many conversations involve talking about nearby objects, such as when commenting on a friend's outfit, or requesting help with a door.

This paper introduces a new approach to improve the speed of AAC input: identifying objects in the user's environment and using information about those objects to generate contextual predictions. Our prototype system, SceneTalk, uses automated image recognition algorithms [8] to identify objects and suggest relevant words and phrases. We present findings from a formative study in which six people with ALS tested the SceneTalk prototype and provided feedback about the usefulness of this approach, as well as their preferences for integrating contextual data into an AAC user

interface. Feedback from this study indicates that using objects in the local environment as contextual predictions may improve the user experience of gaze-based AAC.

RELATED WORK

Slow input speed is one of the most significant barriers to AAC adoption and use for people with ALS [21]. Prior approaches to speeding up AAC input include reducing dwell time [17,20], integrating word and phrase predictions [16], and supporting dwell-free eye-typing [15,22]. However, the theoretical maximum speed of any keyboard-based eye-typing method has been estimated to be only 46 wpm [14], well below the rate of spoken conversation. Alternative text entry methods such as Dasher [24,25] may offer improved performance, but learning an alternative text entry method requires extensive practice.

A common approach to improving AAC speed is to predict likely words using a language model and to suggest them as the user types [23]. Researchers have also explored how the user's context can be used to inform predictions. Higginbotham et al. [11] found that including task-specific predictions in an AAC reduced the number of key presses needed to discuss that subject. Marco Polo [7], TalkAbout [12], and GLAAC [26] offered predictions relevant to the user's location. AACrobat [10] introduced a secondary mobile device application that enabled conversation partners to suggest words and phrases as the AAC user typed. PhotoTalk [1] enabled individuals with aphasia to communicate by sharing photos rather than writing or speaking text. Our work introduces a new approach to generating contextual predictions: identifying nearby objects and suggesting words and phrases related to those objects.

SCENETALK: AN OBJECT-AWARE AAC SYSTEM

To explore the potential of object recognition for AAC word predictions, we developed SceneTalk, a prototype AAC application that identifies objects in the nearby environment and uses those objects to suggest relevant words and phrases.

User Interface

SceneTalk's user interface resembles a traditional character-based AAC system with a QWERTY keyboard (Figure 1). Targets are selected via an adjustable dwell time (defaulting to 500ms). A Play button speaks out the typed text. Word and phrase predictions are shown above the keyboard. A control panel at the top right provides controls for activating the device camera and retrieving contextual predictions, and shows the most recent image captured by the camera.

Interaction Modes

Because context-aware AAC has not been widely tested with users, and because activating the device camera may cause privacy issues if handled incorrectly, we developed three interaction modes for interacting with SceneTalk's contextual predictions. We designed these modes to explore users' preferences for automatic or explicit camera control, and to test whether users preferred contextual predictions to be integrated with linguistic predictions or shown separately. The three interaction modes are:

Always On. The camera captures images automatically at a regular interval. Contextual predictions are shown alongside linguistic predictions, but are presented in a different color and marked with a star symbol (Figure 1).

Click for Suggestions. The user interface is laid out identically to the previous mode. Linguistic predictions are updated as the user types. The user may activate the Camera button to capture an image and load contextual predictions.

Pop-Up Menu. The keyboard prediction row shows only linguistic predictions. The user retrieves contextual predictions by activating the Camera button. Predictions are shown separately via a full-screen overlay (Figure 2).

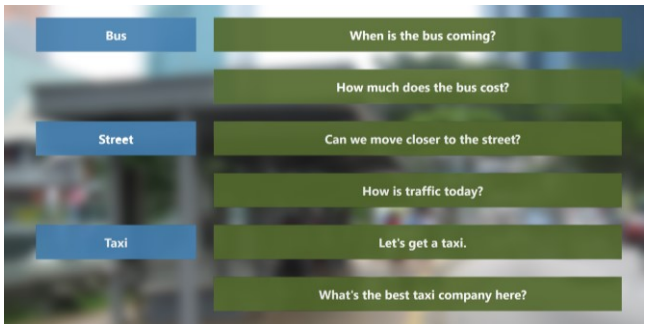

Figure 2. Suggested words (left) and phrases (right) are presented over the camera image in Pop-Up Menu mode.

Recognizing and Labeling Objects

SceneTalk captures images of the environment using the device's rear-facing camera. SceneTalk typically recognizes objects using automated computer vision, but supports crowd-based object labeling as a backup method.

Automated Recognition: SceneTalk uses computer vision to recognize objects in the image. Automatic image captioning has recently been used to provide alternative descriptions of images for blind users [27], but has not previously been used to support communication. SceneTalk uses the Microsoft Cognitive Services API [19], which is based on recent computer vision techniques [8], to identify objects. For each image, the API generates a list of nouns and adjectives describing the scene (e.g., table, café, sunny), along with a confidence value. To identify objects, SceneTalk extracts high-confidence nouns from the list.

Crowd-Based Labeling: SceneTalk also supports crowd-based object labeling using Amazon Mechanical Turk [3]. In this mode, the image is uploaded to a web server, which generates a set of Mechanical Turk tasks. Workers are asked to mark regions in the image corresponding to objects and to provide words or phrases describing each object. Crowd-based recognition may be useful when automated recognition is unavailable, for labeling objects that cannot be identified using current automated methods, and for identifying objects in low-quality images, such as images captured in bright or dim lighting. Crowd workers may also compose relevant phrases based on the identified object, which is not currently supported by automated approaches.

For example, if the user captured an image of a light switch, a crowd worker might suggest the phrase "Can you turn off the lights?" While crowd-based recognition is slower than automated recognition, crowd workers can be used to label images within minutes for a reasonable cost [6], which may still be faster than typing for many gaze-based AAC users, particularly for people with ALS, whose mobility restrictions may result in relatively infrequent changes of scenery.

Word and Phrase Predictions

SceneTalk presents word and phrase predictions as the user types, similar to many other AAC systems (e.g., [23]). SceneTalk's predictions combine linguistic predictions based on the user's prior text with contextual predictions derived from nearby objects.. Linguistic predictions use a word frequency model for determining the initial word and a bigram model for determining subsequent words.

SceneTalk's current interaction mode determines how predictions are displayed, as well as which types of predictions are shown. In *Always On* mode, linguistic and contextual predictions are interleaved and appear as the user types. In *Click for Suggestions* mode, the interface shows only linguistic predictions while typing. Activating the Camera button replaces the linguistic predictions with the top six contextual predictions. In *Pop-Up Menu* mode, only linguistic predictions are presented on the keyboard. Activating the Camera button causes the predictions menu to appear; this full-screen view shows a combination of predicted words and phrases (Figure 2).

SceneTalk can suggest both individual words and complete phrases. Phrases can be generated by crowd workers, but current automated techniques cannot generate relevant phrases. To compensate for this limitation, SceneTalk includes a set of stock phrases that can be attached to any recognized object. These phrases were identified during our formative research with AAC users who have ALS. The stock phrases include "Can you tell me about *X*?", "Can you help me with *X*?", "Can you give me *X*?", and "Can you move *X*?". Phrases are shown when the user begins a new sentence, or when the user selects the *Pop-Up Menu* mode.

FORMATIVE EVALUATION OF CONTEXT-AWARE AAC

Evaluating AAC systems with representative users presents many challenges. Eye-typing for people with ALS can be extremely slow and error-prone [4], making it difficult for participants to provide extended feedback, magnifying the frustration encountered when testing prototypes, and reducing the amount of data that can be collected and analyzed. Furthermore, eye-typing performance can vary significantly throughout the day due to sensor error, user fatigue, and side effects from medication [13]. These challenges make conducting performance evaluations especially difficult for people with ALS. Furthermore, people with ALS typically rely upon their AAC devices for most of their communication, including communicating about physical needs and even life-threatening emergencies. Given the critical importance of maintaining access to communication, we felt that a field deployment of our prototype was beyond the scope of the current work.

For these reasons, we chose to focus on collecting feedback about the usefulness of context-aware AAC, participants' interests and concerns regarding the use of context-aware AAC, and participants' preferences for accessing contextual predictions. We conducted two rounds of evaluation: preliminary interviews based on paper prototypes, and a demonstration and evaluation of our working system.

Paper Prototype Testing

Early in our design process, we collected feedback from six people with ALS (ages 48-54, 1 female). The research team showed participants a paper mockup of SceneTalk, and described how the system could detect and describe objects in the local environment. We asked participants whether the system would be useful to them, and asked them to list situations in which they might use such a system. We also asked participants whether they would prefer to use context-aware AAC predictions via a separate application on their communication device, as a mode in their current AAC device, or completely integrated into their current AAC.

Each of the six participants expressed enthusiasm about our paper prototype: five said that they would use context-aware AAC several times per day, and one said that he would use it several times per week. Participants stated that they would be interested in using context-aware AAC at the doctor's office, supermarket, pharmacy, and around the home. Participants were divided about how contextual information should be integrated into their existing AAC systems: two participants wanted contextual predictions to be integrated into the AAC's keyboard, one preferred to access contextual words and phrases via a separate application, and three said that they would need to try the application themselves before deciding. Feedback from this session was used to guide the design of the SceneTalk prototype and its interaction modes.

Prototype Demonstration and Feedback

After developing a functioning SceneTalk prototype, we presented it to six people with ALS (ages 39-60, 1 female). Three participants used a gaze-controlled AAC keyboard, one typed using a head mouse, and two still used speech as their primary form of communication. Three participants had previously provided feedback about the paper prototype. The SceneTalk prototype was deployed on a Microsoft Surface Pro 3 tablet placed on an adjustable mount. A Tobii EyeX tracker was used to track users' eye gaze.

Participants used the SceneTalk prototype during a single 30-minute session in their home or in our research lab. When possible, participants directly controlled the prototype themselves. However, participants sometimes encountered difficulties using our prototype device due to calibration issues [9], positioning, and fatigue. When tracking became difficult, one of the researchers demonstrated the user interface via the Surface Pro's touch screen. During the session, participants experienced each of SceneTalk's

interaction modes (*Always On; Click for Suggestions; Pop-Up Menu*), presented in random order. Each interface was paired with a randomly-chosen scenario (bus stop, coffee shop, supermarket). For each scenario, pre-generated words and phrases were presented; the words were a subset of those automatically produced by the image recognition API, while the phrases were manually generated by the research team.

After the demonstration, participants answered three Likert-style questions about the prototype (Table 1), indicated their preferred interaction mode, and indicated whether they preferred word predictions, phrase predictions, or a combination of the two. Participants also provided freeform subjective feedback about their experience.

Question	Rating
Overall, how useful were the words and phrases suggested by this application? (1=Not at all useful, 5=Very useful)	3.7 (SD=0.8)
Overall, how likely would you be to use this application? (1=Not at all likely; 5=Very likely)	4 (0.9)
Assuming the camera can be turned off when needed, how do you feel about the application's ability to automatically capture images? (1=Very negative; 5=Very positive).	4.3 (0.5)

Table 1. Summary of Likert-style questions and responses.

Overall, participants rated their experience positively: all recorded Likert scores were 3 or higher. Participants were divided on their preferred interaction mode: two preferred *Always On*, two preferred *Click for Suggestions*, one preferred *Pop-Up Menu*, and one had no preference. Participants were also divided about whether they preferred contextual predictions as words, phrases, or a combination of the two. Four participants preferred an even balance of words and phrases, one preferred mostly phrases with some words, and one preferred only words (no phrases).

DISCUSSION

Because our participants had communication difficulties, they were unable to provide detailed verbal feedback. However, nearly all participants expressed frustration with the speed of their current AAC, and were enthusiastic about being able to more easily discuss items in their environment.

Participants differed in their preferred interface modes and prediction types. This divergence may be due in part to participants' varying abilities, as slower eye-typists may rely more heavily on suggested words and phrases. However, some design features were appreciated by the majority of participants: four of six participants who tested the prototype preferred interaction modes that showed predictions above the keyboard rather than in a separate window (one had no preference), and five of six participants preferred to receive both word and phrase predictions. Given the rapid changes in ability that can occur for people with ALS [4] and the challenges in using gaze-based AAC in certain environments due to ambient light and other factors [13], there may be some benefit in supporting multiple modes of interaction, or adapting the interaction mode to the user's context, such as presenting more predictions if the user appears to be experiencing difficulties using the eye tracker, or weighting contextual predictions more heavily if the user is in an environment in which they are more likely to discuss nearby objects, such as in a store or restaurant.

LIMITATIONS AND FUTURE WORK

A limitation of the present work is that the system has not been evaluated by representative users outside the lab. Testing AAC devices in the field presents many challenges, and these challenges are often amplified for our chosen population of people with ALS. Despite this limitation, we believe that our formative evaluation provides strong support for including contextual information in future AAC devices. Although it was not feasible to replace the AAC devices of people with ALS with a research prototype, publishing information about this novel interface design may bring these concepts to the attention of commercial AAC device makers, providing a route to improvements without requiring users to rely on a prototype AAC system.

While we have conducted this work in collaboration with people with ALS, our approach to providing contextual AAC predictions could be useful to people with a range of abilities. For example, people with aphasia could use SceneTalk to name objects, and foreign language learners could use SceneTalk to practice naming objects in a new language. Adapting SceneTalk to new user groups would likely require changing the user interface: for example, individuals with aphasia may prefer a more visually-oriented user interface, as in SceneTalk's *Pop-Up Menu* mode, rather than a keyboard-based user interface.

An additional area for future work is to improve SceneTalk's capability to generate relevant phrases. Currently, SceneTalk provides phrases from a pre-generated set. Future versions could leverage the crowd or a more robust language model to present more contextually relevant phrases. For example, when identifying a box of tissues, SceneTalk could suggest the phrase "I think I am getting a cold." SceneTalk could also identify groups of objects and infer information about the location (e.g., seeing pots and pans suggests the user may be in a kitchen), and could suggest phrases relevant to multiple detected objects (e.g., "Can you put the book in the bag?").

CONCLUSION

Leveraging information about a user's context to predict words and phrases can help to overcome the slow input speed and high error rate of gaze-based AAC. SceneTalk introduces a new source of contextual data for AAC, that of nearby objects. Our formative evaluation of SceneTalk with people with ALS indicates that objects in the environment can be a valuable source of context, and that these contextual predictions may be integrated into the everyday experience of typing with an eye-gaze based AAC device.

ACKNOWLEDGMENTS

We thank Margaret Mitchell and the Microsoft Research Enable team for their assistance with this project.

REFERENCES

1. Meghan Allen, Joanna McGrenere, and Barbara Purves. 2007. The design and field evaluation of PhotoTalk: A digital image communication application for people. In *Proceedings of ASSETS 2007*, 187–194.
2. ALS Association. Epidemiology of ALS and Suspected Clusters. *ALSA.org*. Retrieved May 25, 2016 from http://www.alsa.org/als-care/resources/publications-videos/factsheets/epidemiology.html
3. Amazon Mechanical Turk. Retrieved January 1, 2017 from http://www.mturk.com/
4. David Beukelman, Susan Fager, and Amy Nordness. 2011. Communication support for people with ALS. *Neurology Research International* 2011.
5. David Beukelman, Pat Mirenda, Kathryn Garrett, and Janice Light. 2012. *Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs, Fourth Edition*. Paul H. Brookes Publishing Co, Baltimore.
6. Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of UIST 2010*, 333–342.
7. Carrie Demmans Epp, Justin Djordjevic, Shimu Wu, Karyn Moffatt, and Ronald M. Baecker. 2012. Towards providing just-in-time vocabulary support for assistive and augmentative communication. In *Proceedings of IUI 2012*, 33–36.
8. Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of CVPR 2015*, 1473–1482.
9. Anna Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In *Proceedings of CHI 2017*, to appear.
10. Alexander Fiannaca, Ann Paradiso, Mira Shah, and Meredith Ringel Morris. 2017. AACRobat: Using mobile devices to lower communication barriers and provide autonomy with gaze-based AAC. In *Proceedings of CSCW 2017*, 683–695.
11. D. Jeffery Higginbotham, Ann M. Bisantz, Michelle Sunm, Kim Adams, and Fen Yik. 2009. The effect of context priming and task type on augmentative communication performance. *Augmentative and Alternative Communication* 25, 1: 19–31.
12. Shaun K. Kane, Barbara Linam-Church, Kyle Althoff, and Denise McCall. 2012. What we talk about: Designing a context-aware communication tool for people with aphasia. In *Proceedings of ASSETS 2012*, 49–56.
13. Shaun K. Kane, Meredith Ringel Morris, Ann Paradiso, and Jon Campbell. 2017. "At times avuncular and cantankerous, with the reflexes of a mongoose": Understanding self-expression through augmentative and alternative communication devices. In Proceedings of *CSCW 2017*, 1166–1179.
14. Per Ola Kristensson and Keith Vertanen. 2012. The potential of dwell-free eye-typing for fast assistive gaze communication. In *Proceedings of ETRA 2012*, 241–244.
15. Andrew Kurauchi, Wenxin Feng, Ajjen Joshi, Carlos Morimoto, and Margrit Betke. 2016. EyeSwipe: Dwell-free text entry using gaze paths. In *Proceedings of CHI 2016*, 1952–1956.
16. I. Scott MacKenzie and Xuang Zhang. 2008. Eye typing using word and letter prediction and a fixation algorithm. In *Proceedings of ETRA 2008*, 55–58.
17. Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast gaze typing with an adjustable dwell time. In *Proceedings of CHI 2009*, 357–360.
18. Päivi Majaranta and Kari-Jouko Räihä. 2002. Twenty years of eye typing: systems and design issues. In *Proceedings of ETRA 2002*, 15–22.
19. Microsoft Cognitive Services. Retrieved January 1, 2017 from http://www.microsoft.com/cognitive-services
20. Martez Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving dwell-based gaze typing with dynamic, cascading dwell times. In *Proceedings of CHI 2017*, to appear.
21. Joan Murphy. 2004. "I prefer contact this close": Perceptions of AAC by people with motor neurone disease and their communication partners. *Augmentative and Alternative Communication* 20, 4: 259–271.
22. Diogo Pedrosa, Maria Da Graça Pimentel, Amy Wright, and Khai N. Truong. 2015. Filteryedping: Design challenges and user performance of dwell-free eye typing. *ACM Transactions on Accessible Computing* 6, 1: 1–37.
23. Keith Trnka, Debra Yarrington, John McCaw, Kathleen F. McCoy, and Christopher Pennington. 2007. The effects of word prediction on communication rate for AAC. In *Proceedings of HLT 2007*, 173–176.
24. Outi Tuisku, Päivi Majaranta, Poika Isokoski, and Kari-Jouko Räihä. 2008. Now Dasher! Dash away!:

Longitudinal study of fast text entry by eye gaze. In *Proceedings of ETRA 2008*, 19–26.

25. David J. Ward and David JC MacKay. 2002. Fast hands-free writing by gaze direction. *Nature* 418.6900: 838.
26. Kristin Williams, Karyn Moffatt, Denise McCall, and Leah Findlater. 2015. Designing conversation cues on a head-worn display to support persons with aphasia. In *Proceedings of CHI 2015*, 231–240.
27. Shaomei Wu, Jeffrey Wieland, Omar Farivar, and Jill Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of CSCW 2017*, 1180–1192.
28. Kathryn M. Yorkston and David R. Beukelman. 1981. Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders* 46, 3: 296–301.